**Introduction**

*“The best way to predict the future is to create it”,* said Abraham Lincoln, without intending, he Formulated accurately the core spirit of each sportsman who tries to predict his future accomplishments. Nevertheless, when I use Machine Learning to analyze many observations of such motivated athletes, it is interesting to identify and measure the importance that certain characteristics have on the success of participating in sports events. Through this rationale and my passion for basketball, I decided to base my ML project on the most famous sport’s tournament for data scientists – the "March Madness".

The National Collegiate Athletic Association (NCAA) Men's Basketball Tournament is informally referred to as "March Madness".

This project is a clear challenge against all odds. there is a 64-team pool with 63 games to predict. Given the sporting nature of a basketball game, it is interesting to identify and measure the importance that certain characteristics influence the success of participating in the tournament. Despite being very difficult to reach great accuracies, people continue to research and try their best. Mathematically speaking, perfectly fill a March Madness bracket is one of the most unlikely things on sport’s events, with a chance of .

The problem is not classification of individual teams, but rather predicting the outcome of a match between any two teams. Since I have the data of the last 25 years, I would try to use Machine Learning to find out what statistics most correlate with a team winning a match-up.

My ML model will analyze information about two teams (Team 1 and Team 2) as input, and then output a probability of Team 1 winning that matchup.

By running this model over the first-round matches (that I set as the input), I can simulate the whole tournament until the final, and predict the big winner.

In order to fill out tournament brackets with high predictive accuracy, many computer simulations and algorithms have been developed to model the tournament and attempt to explore the effective strategies for March Madness prediction. By reviewing the most successful models (the winners of Kaggle’s previous years contests), I can define three major factors that are strongly influences the prediction accuracy - teams’ seeds, rating of teams learned from season games and team’s possession. On top of those factors I’ll add more variables from the Kaggle’s “March madness” data set and I’ll examine the possibility for external data.

On a personal note, beside dealing with one of my favorite sports, I hope to get more added value from the methods of this project that will be used by me on my daily work. Since I have a lot of performance data, and this kind of a performance-oriented prediction, should be beneficial for this goal.

**DATA**

Kaggle’s March Madness is one of the famous Data Science contest in the world for several years. Because of the contest seniority and popularity, the dataset is very neat and clean from technical biases and Reasonable need for validations.

As part of efforts to strengthen the prediction, I searched for external sources of data Beyond Kaggle’s Datasets.

Finally, I focused on two options that based on research over the net and by guts feeling –basketball team budget and traveling distance.

For The first one, I looked for data and found the number of 2016, that was correlated with 2017 tournament results. Of course, it wasn’t wide enough for inputted the rest of the observations.

About the other external data option, I found a perfect data set that was made by one of the previous year’s contestant. This Dataset contain all the coordinate of the tournament neutral courts, and the location of each team home court. By this Dataset, I was able to build a path for each team and calculate by GoogleMaps API, the exact travel distance before each phase.

The structure for my data set is very simple and based on 4 internal (Kaggle’s Datasets) and one external dataset, as detailed below.

External Dataset

Internal Datasets

(Kaggle’s)

NCAATourneySeeds

NCAATourneyCompactResults

Coordinate of the tournament games location since 1985

Calculate the Traveling distance for each team in each contest round

RegularSeasonDetailedResults

Final CSV

Teams

**Time Frames**

The structure of the model for predicting the tournament based on each year regular season and the relevant data that we can get before the first round.

Based on it I build the interval as yearly one, with no influences beside the specific year effects.

In addition, following the eternal changes in sports, and specifically the athletic level that got improved on the last serval years. The model test must contain a mix between the latest years and the old ones, so the training will catch all the latest trends and still let us check the model on more relevant observations.

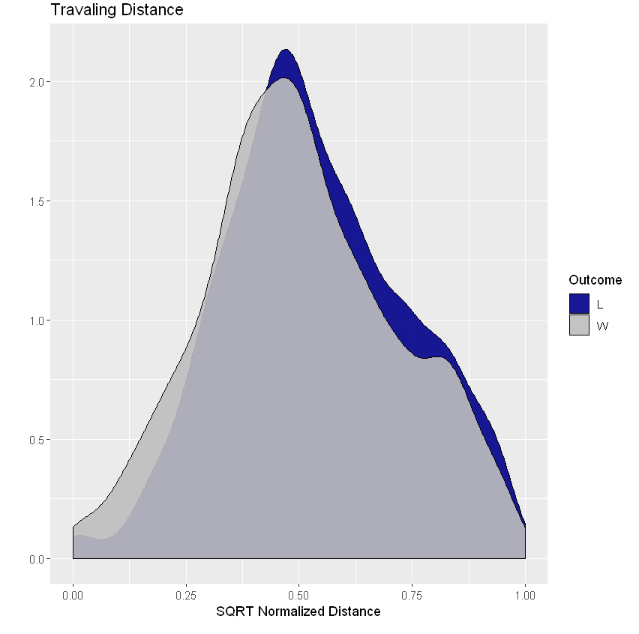
**Outcome**

My outcome will be the winning team and the loosing one from each NCAA march madness tournament match.

**DATA Exploration**

The strategy of the data exploration was to find the main factors, allegedly, for winning a tournament match.

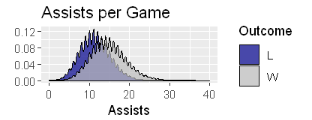
**In this section I’ll describe the main variables that I found efficient for the model prediction:**



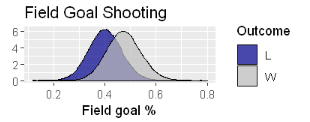
This Graph shows the external data that I described on the previous section. As we can see there is slightly advantage for teams who have shorter journeys between the tournament match-ups. Even when this visualization is not as signigican as the rest, I’ll use it in my model.



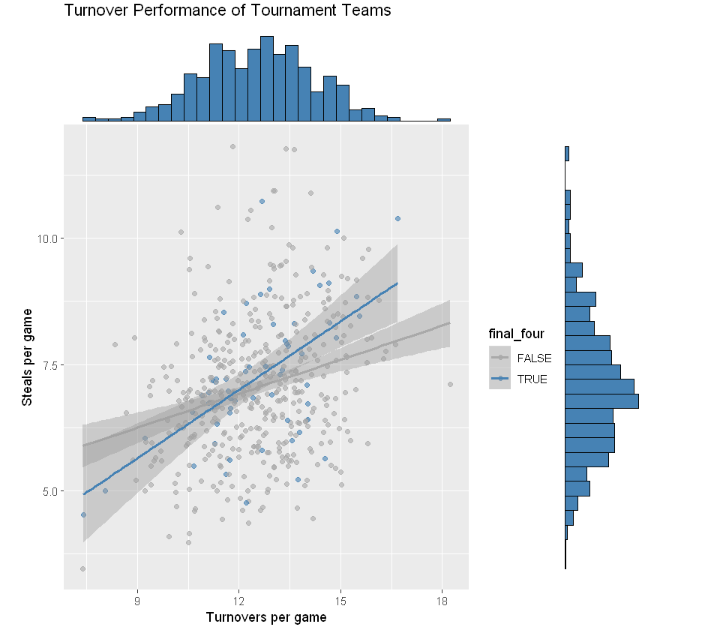
Not surprisingly, The Teams seed’s is highly correlated with the number of tournaments winning. This basic label must to be a part of our module’s data.



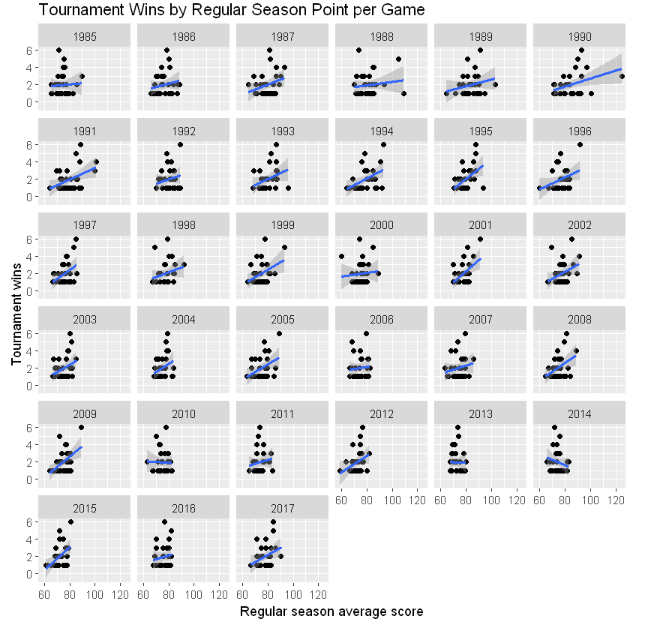
The state of the assists is curvy because of the roundly numbers. in comparison to other stats, it’s more significantly correlated to the winning teams. That’s why I’ll use it in my prediction.



From the density plots, it appears that Final Four teams do shoot better from the floor during the regular season. For this reason, I’ll would implant this regular season state in the prediction model.



The ratio of steals to turnovers is positive for all tournament teams, however the relationship appears to be stronger for Final Four teams indicating that this ratio be be a good predictor of tournament success.



As part of my analysis, I tried to figure out what is the best way to use the regular season scores. I wondered between the volume of the points and the margin between the winners to the loser’s points, as the high causality between those two made me to choose only one of them. After comparing them by visualization, I decided to choose the score rate with an higher positive trend.

**Data Enrichment**

All the regular season’s stats were manipulated to be an annually average for each team. This way the relations effects only on the specific year.

**Missing values**

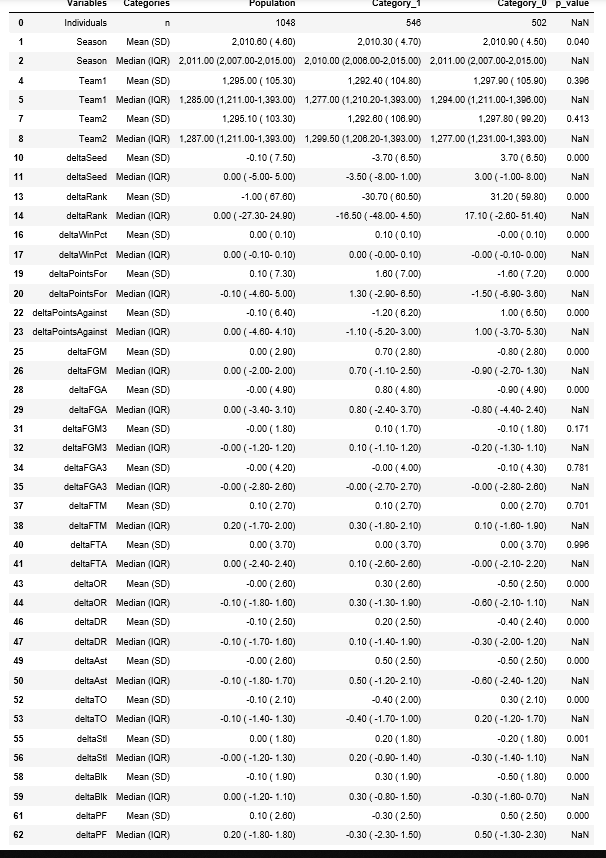
As mentioned, the Kaggle’s Datasets were edited many times before my usage, so I had only little few missing values.

**Outliers**

Will be completed.

Due to this current Deadline I didn’t manage to finish the treat of outliers.

**Data Retrieval**



## **Models**

Because of the minor amount of observations, I divided the data in proportions 80:10:10 Train:Test:Dev.

I checked the quality of it by the Table1 feature of the “Mahchkar” package. The only column that wasn’t completely at random according to P-value check between the data set, was removed.

Because my model is very specific – predicting the winner of basketball games, regression is the only way to base the model on.

My way to evaluate my models is the sum of squared errors and the proportions of the accuracy. I chose those simple measurements, because I tried to get the best platform for my model to be compared to other predictive models for the march madness. Those two are uses by most of the Kaggle’s components

**Conclusion**

The experience of building the first End-to-End model was significant for the roots of my perspective as a junior data scientist. I tried, as hard as I could, to understand the statistic behind the models. This priority made me to invest most of the time on videos and articles. Although it was On account the technical issues, but at the end of the process, I’m glad that two of my main insights, came from the statistical point of view. The first one was the structure of the data-set timeline, that was mixed up and didn’t have any benches of yearly tournament. This decision significantly improved my predictions. Furthermore, delve into the random forest theory, what eventually made me change part of my data structure, and improve my model, that random forest was the best prediction for it.

I spend so much time in this project and I have a lot of confidence and enthusiasm towardnext project.

Meanwhile I can follow 2019 march madness with an added value of estimating my model.

I ran some code that build up the tournament bracket according to my predictions and as of the day I wrote those lines I have a Log Loss of 0.48 (top 5 score for 2018 , but it’s only the beginning):

